Szintaktikai elemzés szerepe a biológiai eseménykinyerés kulcsszavainak detektálásában

Móra György[1], Molnár Zsolt[2], Farkas Richárd[3]

[1] SZTE, Informatikai Tanszékcsoport,
H-6720 Szeged, Árpád tér 2.
gymora@inf.u-szeged.hu
[2] Acheuron Hungary, Kemo- és Bioinformatikai Csoport,
H-6720 Szeged Tiszavirág u. 11.
zsoltm@acheuron.hu
[3] SZTE, MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
H-6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: Angol nyelvű élettudományi cikkekben szereplő biológiai események kulcsszavainak meghatározásához felhasznált hat nyelvi elemzőt hasonlítottunk össze. Biológiai esemény a szövegben leírt biológiai tény, folyamat. Az esemény kulcskifejezése az eseményt indukáló szövegrész, argumentumai a szövegben található biológiai entitások, mint például fehérjék, gének. A szövegből nyert statisztikai és nyelvi jellemzők felhasználásával döntési fa és szupport vektor gépi osztályozókat tanítottunk. A modellek teljesítménye közvetve információval szolgál az adott nyelvi elemző felhasználhatóságáról a kulcsszókinyerési feladaton.

1 Bevezetés

A tudományos publikációkban rejlő hasznos információk megszerzése sokszor komoly problémát jelent az információáradattal küzdő kutató számára. A biológia és az élettudományok területén ezért egyre nagyobb igény mutatkozik olyan információkinyerő rendszerekre, amelyek a publikációkból (szabadalmak, újságcikkek, konferenciakiadványok) tényeket, adatokat nyernek ki kereshető, strukturált formában. Az elmúlt években az érdeklődés fókusza az interaktáló fehérjepárok azonosításáról az összetettebb, részletesebb adatok kinyerésére tevődött át.

Az ún. *biológiai események* nem csak kétszereplősek lehetnek, egy vagy akár több fehérje is szerepelhet egy eseményben. Emellett az események más eseményekre is hivatkozhatnak, komplexebb tudásbázist létrehozva. A biológiai események jóval pontosabb adatokat tartalmaznak a biokémiai, sejtbiológiai folyamatokról, történésekről, mint a fehérje-interakciók, így értékesebb, piacképesebb adatbázisokat lehet építeni belőlük.

A gépi tanuláson alapuló eseménykinyerő rendszerek fejlesztése a GENIA Event Corpus [2] megjelenéséhez köthető, amely az első komplex biológiai eseményeket tartalmazó manuálisan annotált korpusz. A BioNLP2009 Shared Task on Event

Extraction elnevezésű eseménykinyerési verseny [1] volt az első, amely ezt a problémát tűzte ki feladatául.

Egy teljes biológiai esemény egy, az eseményt indukáló kulcskifejezésből, a résztvevő entitásokból és az őket összekötő esemény típusából áll. Az entitások fehérjék, gének és egyéb molekulák nevei. A versenyen ezen entitások szövegbeli előfordulásait ismertnek tekintették. A résztvevő rendszerek túlnyomó többsége két részfeladatra bontotta a problémát. Első lépésben az indukáló kulcsszavakat azonosították, majd szereplőket rendeltek ezekhez az entitáshalmazból. A verseny tapasztalatai, valamint a legjobban teljesítő rendszerek eredményei alapján megállapítható, hogy a függőségi és más szintaktikai elemzők kimenetéből nyert jellemzők jelentősen javítják a gépi tanulási rendszerek teljesítményét.

Cikkünkben a kulcskifejezés azonosításának részproblémájára koncentrálunk és négy különböző szintaktikai elemző kimenetének felhasználásával nyert jellemzőkészletet hasonlítunk össze. A gépi tanuló modellek teljesítményét értékelve a verseny által biztosított adathalmazokon, megállapítható, mely elemzők vagy melyek kombinációja adja a legjobb eredményt, illetve tárgyaljuk az egyes elemzők (melyeknek elméleti alapjai is különböznek) előnyeit, hátrányait, alkalmazhatóságuknak feltételeit. A jellemzőkészlet a szintaktikai és függőségi elemzők eredményein kívül a szavak más egyéb tulajdonságait is tartalmazza, ám ezek minden elemző esetében megegyeznek. A jellemzőkészlet mintájául a BioNLP2009 Shared Task on Event Extraction verseny első helyezettjének kulcskifejezés jelölő rendszere szolgált [3].

A különböző elemzőknek a feladaton elért eredményeit összehasonlítva láthatóak azok előnyei, illetve hátrányai a kulcsszódetektálásban. Az eltérő nyelvészeti megközelítések különböző összefüggések kinyerésére alkalmasak, ezért is fontos a feladatnak megfelelő kiválasztása.

2 Nyelvi elemzők

A vizsgált nyelvi elemzők két csoportba sorolhatóak. A függőségi elemzők (*dependency parser*ek) a mondat szavai közötti kapcsolatokat függőségi fa formájában ábrázolják. A fa minden pontjához egy szót rendelnek – amelyeknek pontosan egy őse van –, kivéve a virtuális gyökérelemet. A pontok és őseik közötti élek, valamint ezek címkéi definiálják egy mondat szerkezetét. Szabad szórendű nyelvek elemzésére különösen alkalmas, lévén a fa szerkezete a szavak sorrendjétől nem, csak a közöttük lévő nyelvi kapcsolattól függ.

A másik csoportba a frázisstrukturált nyelvtant használó elemzők tartoznak, amelyek a mondatokat hierarchikus formában, konstituensfaként írják le. A csomópontok igei, főnévi, stb. nyelvi csoportokat jelentenek, a fa gyökerében a mondatot reprezentáló pont van. Két egymás melletti csoport alkothat egy magasabb szintű csoportot, így a szavak sorrendjétől is függ, mely szavak képezhetnek egy csoportot. A fa pontjai nem a mondat szavainak felelnek meg, mint a függőségi fánál, hanem a mondatot alkotó hierarchikus szerkezeteket jelölik. A függőségi formátumtól eltérően itt a pontok címkéi tartalmazzák a felhasználandó információt, az élek az egyes csoportok elemeit, azok felbontását adják meg hierarchikus formában.

A *PCFG (Probabilistic Context-Free Grammar)* elemzők környezetfüggetlen nyelvtan segítségével elemzik a mondatokat. Az egyes csoportok valószínűségeit kombinálva határozzák meg a szöveg legvalószínűbb konstituens elemzését.

A *HPSG (Head-driven phrase structure grammar)* elemzők összetett, struktúrált "szótárak" és szabályok alapján építik fel a frázisok hierarchiáját. Minden frázisnak van egy feje, amely kitüntetett szerepű a kifejezés felépítésében. A szavak és frázisok tulajdonságait egymásba ágyazódó hierarchikus kulcs-érték párok adják meg. Ez a frázisstruktúra felbontható a beágyazások mentén, és faszerkezetben ábrázolható.

A cikkben felhasznált nyelvi elemzők:

3. ***Bikel:*** *Mike Collins* függőségi elemzőjének *Dan Bikel* által implementált változata
4. ***CCG:*** A *C&C Tools* függőségi elemzője biológiai doménre
5. ***Enju:*** Valószínűségi *HPSG* modellt használó szintaktikai elemző. Akár több lehetséges elemzési kimenetet is generál a valószínűségeik sorrendjében. A felhasznált változatot a *GENIA* korpuszon tanították.
6. ***Gdep:*** A *KsDep* függőségi elemző *GENIA* korpuszon újratanított változata
7. ***McClosky-Charniak:*** *Charniak és Johnson statisztikai* elemzőjének *David McClosky* által továbbfejlesztett biológiai doménre adaptált öntanulást alkalmazó változata
8. ***Stanford:*** *PCFG* elemző, frázisstruktúrált és függőségi formájú kimenettel.

3 Kulcsszavak detektálása

A biológiai események az élettudományi cikkekben szereplő valamilyen biológiai tényt vagy folyamatot írnak le. Az eseményeket jelző szövegrészlet az esemény kulcskifejezése. Az egyszerű statisztikai modellek helyett a nyelvtani elemzőkkel előállított, a szavak mondatban betöltött szerepét leíró jellemzők használata válik elterjedtté [1]. A kulcsszavak meghatározása osztályozási feladatként, gépi tanulási módszerek segítségével történt. A tanítóadatbázis a versenyen kiadott *train* halmaz volt, míg a kiértékelést a *development* halmazon végeztünk. Az *infogain* alapján leszűrt kétezer legjobb jellemzőn tanított *C4.5* döntési fa (*Weka J48*) és az összes jellemző felhasználásával tanított szupport vektor modellek (*libsvm*) eredményeit mértük meg.

A jellemzőkészlet mintájául a BioNLP2009 SharedTask on Event Extraction versenyen legjobban szereplő rendszer kulcsszódetektáló rendszere szolgált. A jellemzők három nagyobb csoportra oszthatók:

- **Token jellemzők:** A mondatok szavakra bontását a GeniaTagger tokenizálójával végeztük. A jellemzőkészlet tartalmazta a szavak gyökerét, amit a *Porter stemmer* állított elő, a szavak karakterenként vett bi- és trigramjait.

- **Numerikus jellemzők:** Ezek a jellemzők a szó adott tokenszámú környezetében és a mondatban található biológiai entitások számát, a mondatban található egyedi szavak számát adják meg.
- **Nyelvi jellemzők:** A nyelvi jellemzőket a függőségi fa az adott szóból kiinduló 1-3 mélységű útvonalai és az útvonalak végén található szavak mondatbeli funkciói alkották. A frázisstruktúrált elemzők kimenetét a függőségihez hasonló formában használtuk fel. Az *Enju* kivételével az összes ilyen elemző kimenete rendelkezésre állt *Stanford függőségi formátum*ra alakítva.
- **Szomszédos szavak:** Minden szóhoz a nyelvi fában a szülő szó, a gyerek szavak, illetve a szavak közvetlen környezetében található tokenek összes tokenjellemzőjét hozzárendeltük.

4 Eredmények

Jelen munkában a különböző nyelvi elemzők használhatóságát az általuk előállított jellemzők felhasználásával tanított kulcsszódetektáló modellek eredményeivel jellemezzük (1. táblázat). A C 4.5 modell kiértékelését keresztvalidációval is elvégeztük. Az elemzők nagy része biológiai doménre készült, de vannak közöttük általános szövegen tanítottak is. A kis eltéréseket és a magas pontosságot az okozza, hogy a szavak csak kis aránya kulcsszó, így a "nem kulcsszó" osztály előfordulása magas.

A keresztvalidáció során az elemzők nem mutattak jelentős eltérést, de a *dvelopment set*-en a *Stanford parser* teljesített legjobban a döntési fa modellel, a tanító adatbázis relatív méretének csökkenésével javult a teljesítménye.

1. táblázat: A különböző nyelvi elemzők teljesítménye a kulcsszó-meghatározási feladaton.

	Bikel	**CCG**	**Enju**	**GDep**	**M-C**	**Stanford**
C 4.5	96,696	96,660	96,655	96,783	96,681	**96,925**
C 4.5 k.v.	97,618	97,638	97,583	97,635	97,645	97,617
libSvm	96,730	**96,804**	96,450	96,552	96,635	96,408

Köszönetnyilvánítás

A kutatást – részben – a BAROSS_DA07-DA_TECH_07-2008-0028 projekt támogatta.

Hivatkozások

1. Kim, J-D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 Shared Task on Event Extractionin Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop (2009)
2. Kim, J., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature, BMC Bioinformatics (vol. 9) (2008)
3. Bjorne, J., Heimonen, J., Ginter, F. Airola, A., Pahikkala, T., Salakoski, T.: Extracting Complex Biological Events with Rich Graph-Based Feature Sets, BioNLP2009 Workshop Companion Volume for Shared Task Association for Computational Linguistics (2009)